A SUMMARY OF CORRELATION METHODS

H. W. Stoker
Florida State University

Introduction

Whenever a situation presents itself in which a pair of scores is obtained for each member of a group of individuals it seems that someone "runs a correlation". More likely than not the person will compute a Pearson r.

The purpose of this paper is twofold: first, to develop the concept of correlation itself and second, to describe and illustrate several kinds of correlation techniques and to list the assumptions which must be met to apply justifiably each kind of correlation technique. No attempt will be made to include any formulae for computing particular coefficients of correlation. References will be provided at the end of the paper to satisfy this need.

The term "correlation" is used in a broad sense to indicate the type of change, increase or decrease, in one variable as a second variable takes on different forms or values: e.g. "correlated curricula", "correlated services," "correlated textbooks," etc. The term "correlation" as used in statistics (and the term should be called "statistical correlation") implies a well defined mathematical function of measurements on the two variables whose behavior is consistent with what is implied to the term "correlation" used in the above examples.

People have different ideas when they speak about correlation and hence, as we might expect, there are many types of statistical correlations. There is a possibility that given a specific situation no correlation technique would satisfy all the requirements and differences of opinion which arise in the selection of one of the many correlations that can possibly be used.

In a typical problem to be analyzed through correlation there are two variables of interest and measurements are available on both variables for each individual or instance in a group. For example, suppose that achievement and ability in arithmetic for a group of students in a certain curriculum are measured in terms of scores on an achievement test and scores on an ability test. We may be interested in studying or evaluating the way in which the behavior on one variable is accounted for by behavior on the other. If we

can control one variable then we may be able to control the other, at least to some extent. In the case of the achievement and ability scores in arithmetic, if we find that achievement scores increase when ability scores increase we may be able to produce students with high achievement by increasing ability scores of future students (perhaps through an adjustment in course work). If one variable is measured more easily than the other the concept of correlation can be of help in predicting the value on the variable which is more difficult to measure. For example, if we study the correlation between school enrollment and total expenditurewe will, in all probability, find that higher school enrollment leads to higher expenditure. A study of correlation should be of help in predicting the amount of funds which will be necessary for increasing numbers of students.

After we have decided to study the correlation between two variables our first step is to collect measurements for each individual on the two variables under consideration. In a situation where it is possible to measure the variables in more than one way we should select those measurements which reflect most clearly the significance of the two variables. A measure of correlation should be sought that will give the most accurate description of the concept of correlation. If a complicated correlation formula is avoided in favor of a simple one, from the standpoint of mathematical computation, careful consideration must be given to the answers obtained before arriving at final conclusions or making recommendations.

General Properties of Correlation

Statistical correlation has the following properties:

1. "correlation" is positive when large values of one variable are associated with large values of the other;
2. "correlation" is negative when large values of one variable are associated with small values of the other; and
3. "correlation" is small or close to zero when neither large nor small values of one variable are associated consistently with values of the other.

The scale used in indicating correlation runs from -1 to +1 with zero as the mid-point. As the correlation coefficient approaches either -1 or +1 we have a "high" relationship (high negative or high positive). As the correlation coefficient approaches zero we say we have a "low" relationship.

Descriptions of Correlation Coefficients

Pearson r

The most widely used measure of correlation is the product moment correlation developed by Karl Pearson about 1900. The Pearson product moment correlation is generally accepted to be the "best" correlation method to be used. The use of this technique requires that the two variables have a bivariate normal distribution. This is equivalent to the two conditions (1) linear regression between the two variables (i.e. the scores when plotted on a pair of axes tend to fall along a straight line) and (2), for any given value on one variable the distribution of scores on the other variable is normal and these normal distributions have the same variance. The property of same variance is also referred to as homoscedasticity. The assumptions for computing the Pearson r will probably be met if the two distributions are unimodal and fairly symmetrical.

Example: A principal is interested in examining the relationship between scores on an ability test and scores on an achievement test for a group of high school freshmen. He plots the pair of scores for each student on a pair of axes (the resulting plot is called a scattergram). The scores on both tests constitute continuous scales and the linearity of relationship and homoscedasticity can be checked by observation of the scattergram. If these last two conditions are apparent then all assumptions have been met and one may compute the Pearson r.

Rank Difference Correlation

Frequently variables cannot be measured in absolute units or the absolute units that may be developed are subject to arbitrary definitions (e.g., patriotism, social adjustment, etc.). For example, a salesman could be ranked from highest to lowest in terms of personal appearance, students could be ranked in order of adjustment, etc. A rank difference correlation (usually referred to as Spearman's rho--ρ) enables one to determine relationship between variables of this type. The computation ρ is relatively simple and the method is occasionally used in small samples (less than 30 pairs of scores) as a substitute for r. When used as a substitute for r it is possible to obtain a good, though approximate, estimate of r through the use of established formulas, if the assumptions for r as described above are justified. However, inasmuch as these assumptions are difficult to meet when dealing with rank-ordered variables, the calculation of rank order coefficients of correlation is usually restricted to the measurement of relationships between rank orders.

Example: In a class of beginning speech students the teacher ranks the students on "stage presence" and "knowledge about subject" in an attempt to determine the relationship between these variables. Both variables are expressed in terms of rank rather than in terms of a continuous scale, hence, ρ would be a justifiable correlation coefficient to compute in this situation.

Biserial r

In some situations one variable is measured continuously but the other is measured in terms of a two-point scale. There may be reason to believe that the two-point scale is actually a two-way splitting of a continuous scale. For example, whether or not a student passes a course will depend on his scores on tests over a period of a semester. The final scale on which the pass or fail judgment is made is a continuous one although the actual measurement is in terms of the two-way split, pass or fail. In situations of this type, the biserial r (r_b) enables one to estimate the degree of relationship between the variables. The calculation of r_b requires that the scales for both variables be continuous, that homoscedasticity exists, and that the distributions of both underlying variables are normal.

Example: An estimate of the relationship between test scores and graduation from high school is desired. Test scores represent a continuous scale; the decision as to whether or not a student graduates is based on a continuous scale. The distribution of students along this scale as well as along the scale of test scores could be assumed to be normal. The r_b would therefore enable one to estimate the relationship between the variables.

Under certain conditions the biserial correlation provides a good estimate of the Pearson r. However, when used in this manner, the assumptions for the calculation for the Pearson r must be satisfied.

Point Biserial

If one of the variables of interest is measured on a continuous scale and the other is a true dichotomy (e.g., male - female, living - dead). The point biserial coefficient of correlation (r_{pb}) may serve to estimate the relationship between the variables. The r_{pb} is frequently used as a measure of the degree to which the continuous variable discriminates between the two categories of the dichotomous variable. To compute r_{pb} we need only to assume homoscedasticity.

Example: In a physical education class the instructor wishes to determine if there is a relationship between sex and strength of grip. The first variable is a true dichotomy and the second can be measured on a continuous scale. Assuming homoscedasticity we can then compute r_{pb}.

Inasmuch as the point biserial coefficient is not restricted to normal distribution for the dichotomized variable it is generally considered to be more widely applicable than is the biserial r. Only under special circumstances can r_{pb} be used as an estimate of Pearson r.

Tetrachoric r

If both variables are measured on a continuous scale but both have been dichotomized then the tetrachoric correlation coefficient (r_t) could serve to indicate the degree of relationship between the variables. Calculation of r_t requires that both variables have been measured on a continuous scale and that the assumptions of linearity and normal distribution of the underlying variables exists. It is recommended that r_t be used only in situations where the number of pairs of scores is at least 200 and preferably more than 300. Special care should be exercised to avoid the computation of r_t in situations where the split in either variable is extremely one sided (90%-10%, 95%-05%, etc.).

Example: Three hundred students have taken a test and the experimenter is interested in the relationship between their test scores and their IQ's. The students are divided into two groups, those with IQ's of 100 or more and those with IQ's below 100. The students are also divided into two groups based on the test scores--those with high scores and those with low scores (i.e., the total score distribution is divided into upper and lower halves). Assuming linearity of relationship and normality of the underlying score distributions we may compute the tetrachoric correlation coefficient.

Two common correlations techniques exists which place no restriction on the user in terms of assumptions with respect to the distributions of the variables. These techniques fall into a class of statistics known as nonparametrics. Both techniques are closely related to the Chi-square distribution. This type of distribution is based on frequencies rather than measurements.

Coefficient of Contingency

When the two variables are classified in categories, the coefficient of contingency (C) may be used. C is directly related to the chi square technique which can be used

to test the independence of two variables. One of the major advantages of C lies in the fact that no assumption pertaining to the distributions of the variables need be made. Computation of C requires only nominal measurement, the least refined type of measurement we can have. The maximum value of C depends on the number of categories used in the contingency table and approaches 1 only as the number of categories for the variables increases beyond 10. As an example of this limitation on the maximum C consider a contingency table where both variables have 3 categories (e.g., color of hair: blonde, brown, red, and color of eyes: blue, brown, green). The maximum possible value for C in this example is approximately .82.

Example: An indication of the relationship between socio-economic class (high, middle, and low) and attitude toward new taxes (approve, undecided, and disapprove) is desired. No assumptions regarding continuity or distribution are necessary and hence the coefficient of contingency is a justifiable computational method in this example.

The contingency coefficient is not directly comparable to any other measure of correlation.

Phi Coefficient

If the variables of interest are true dichotomies (male - female, living - dead, etc.), the phi coefficient (ϕ) may be applicable. In actual practice ϕ is used when the variables are not true dichotomies. It is recommended that this technique not be used when the split on either dichotomy is extremely unequal (90%-10%, 95%-05%, etc.) because of the influence of the split on the maximum attainable value of ϕ.

Example: Two items on a test are scored "pass" or "fail". The phi coefficient is a justifiable method of estimating the relationship between performance on the two items.

Other Correlations

Three other correlational techniques should be mentioned. These are partial correlation, multiple correlation, and curvilinear correlation.

Partial correlation involves the removal of the effects of a third variable (or number of other variables) from the two variables of interest. It would, for example, allow for the investigation of the relationship between reading test scores and intelligence test scores for a group of children of different ages. Through the use of partial correlation one could remove the effect of age differences on reading ability.

Multiple correlation is a method for examining the relationship between several variables at the same time. The method has great value in the problem of the prediction of one variable from several variables. The coefficient of multiple correlation (R) presents an estimate of the relationship between the one variable and several others on which it appears to depend. The multiple R is an indication of how accurately the scores from these several independent variables represent the actual value of the dependent variable.

Curvilinear correlation techniques should be employed when the assumption of linearity, needed in all other methods, is not apparent or even suggested. This will occur when one variable increases at a much more rapid rate than the other or in a situation where as one variable increases, the other increases up to a point then decreases. The technique here involves the use of the correlation ratio or eta (η) coefficient. The variables involved must be continuous.

Final Comments

Finally, with regard to the interpretation of any correlation coefficient, it is important to remember that correlation is always relative to the conditions under which it is obtained. One must constantly keep in mind the variables being measured, the way in which they are measured and the choice of correlational method employed. We should not use the term, "correlation" as if it were an exact quantity: in reality it is completely relative and must be interpreted in light of the circumstances involved.

Summary of Correlational Methods

Name	Symbol	Variables	Assumptions
Pearson Product-Moment	r	Two continuous variables.	Linear relationship between variables. Homoscedasticity.
Rank-difference Method	ρ (Rho)	Two rank ordered variables.	
Biserial Correlation	r_b or r_{bis}	Two continuous variables, one has been dichotomized.	Homoscedasticity. Both variables normally distributed in the population.
Tetrachoric Correlation	r_t	Two continuous variables - both have been dichotomized.	Homoscedasticity. Normality of distribution in population for both variables. Use with minimum of 200-300 pairs in sample.
Point-biserial Correlation	r_{pb}	One variable continuous - one variable a dichotomy.	Homoscedasticity.
Phi-coefficient	ϕ	Both variables true dichotomies.	
Contingency Coefficient	C	Both variables can be placed into two or more nominal categories.	
Correlation Ratio or Curvilinear Correlation	η	One variable continuous and the other continuous or discrete.	
Partial Correlation	$r_{xy.z}$	All continuous	Same as for Pearson Product-Moment.
Multiple Correlation	$R_{x.yz}$	All continuous	Same as for Pearson Product-Moment.

References

1. Ferguson, George A. Statistical Analysis in Psychology and Education. New York: McGraw-Hill Book Company, Inc., 1959.

2. Guilford, J. P. Fundamental Statistics in Psychology and Education. New York: McGraw-Hill Book Company, Inc., 1965.

3. Kendall, M. G. Rank Correlation Methods. London: Charles Griffin & Company, Limited, 1948.

4. Siegel, Sidney. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill Book Company, Inc., 1956.